様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ６ 日現在

機関番号：１５４０１
研究種目：挑戦的萌芽研究
研究期間：2013～2015
課題番号：２５５４００１２

研究課題名（和文）情報量規準最小化に基づくモデル選択法の理論的考察

研究課題名（英文）Theoretical considerations of model selection method based on a minimization of an information criterion

研究代表者
柳原　宏和（Yanagihara, Hirokazu）

広島大学・理学（系）研究科（研究院）・准教授

研究者番号：７０３４２６１５

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：本研究課題は，情報量規準最小化に基づくモデル選択法に関する研究である．モデル選択に用いる情報量規準として多くの規準量が提案されている．モデル選択に使用する情報量規準の特性を標本数のみを無限大とする漸近理論である大標本漸近理論と標本数だけでなく目的変数ベクトルの次元数も無限大とする漸近理論である高次元大標本漸近理論により評価する．その結果から，実解析において，どの情報量規準を使用すればよいかということに関する判断材料を提供する．

研究成果の概要（英文）：In this research, we study model selection method based on a minimization of an information criterion. There are many information criteria. We evaluate theoretical properties of an information criterion used for model selection by a large sample asymptotic theory such that the sample size goes to infinity and a high-dimensional and large sample asymptotic theory such that the sample size and the dimension of a vector of response variables go to infinity simultaneously. From obtained results, we provide standards of judgment with regard to deciding an information criterion.

研究分野：統計科学

キーワード：モデル選択　情報量規準　多変量線形回帰モデル　高次元データ

１．研究開始当初の背景

統計解析において，どの変数を用いるかにより想定する統計モデルが異なってしまい，使用する統計モデルにより得られる結果が大きく間違ってしまう可能性がある．考えうる候補のモデル群から最適なモデルを選ぶ手法をモデル選択手法と言い，その需要の高さから，古くから多くの統計学者によりモデル選択法に関する研究が多く行われている．モデル選択法の中でも，各候補のモデルで Akaike (1978) により提案された赤池情報量規準 (Akaike's Information Criterion: AIC) や Schwarz (1978) により提案されたベイズ型情報量規準 (Bayesian Information Criterion: BIC) に代表される情報量規準を計算し，それを最小にするモデルを最適なモデルとして選ぶ，情報量規準最小化に基づくモデル選択法が様々な応用分野で使用されている．AIC と BIC は，候補のモデルの最大対数尤度の−2 倍にモデルの複雑さに対する罰則項を加えることにより定義される．その形の簡便さから，AIC や BIC はモデル選択規準最小化に基づくモデル選択法において，最も多く用いられる情報量規準である．2 つの情報量規準は形こそ罰則項の違いだけでよく似たものではあるが，AIC は予測の精度を向上させることを，BIC は真のモデルを選ぶ頻度を上げることを目的とし，まったく異なった目標を達成させるために使用されることが通例である．AIC や BIC ではモデルの良さをカルバック=ライブラ (Kullback-Leibler: KL) の距離で測るが，AIC を使用したからと言って，選ばれたモデルの予測 KL 距離である予測誤差が小さくなるという理論的保証はほとんど無い．また，真のモデルが含まれているパラメトリックモデル群において，BIC を使用したモデル選択では，標本数 n を無限大としたときに真のモデルが選ばれる確率が 1 に収束するのに対し，AIC では 1 に収束しないことがよく知られている．ところが，特定の多変量モデルにおいて，観測値の次元数 p も n と共に無限大としたら，まったく逆の結果を得ることもある．このように，どの情報量規準を使用すれば良いかという問題は重要かつ深刻な問題であるにも係らず，どの情報量規準を使用するかという判断基準があいまいな状態である．

２．研究の目的

情報量規準最小化に基づくモデル選択では，

(I) 使用する統計モデルの予測の精度を上げたいときは AIC を使用し，真のモデルを選ぶ頻度を上げたいときは BIC を使用する．

(II) 使用する統計モデルの予測の精度をより上げたいときは，AIC が持つ候補のモデルの予測 KL 距離に対するバイアスが小さくなるようなバイアス補正 AIC を使用する．

(III) どのような場合でも BIC は真のモデルを最適なモデルとして選択する確率が標本数 n の増加に伴い 1 に収束するが，AIC は 1 に収束しない．

といったことが通例である．その通例に従って情報量規準の使い分けが行われてきたが，実はその理論的保証は (I) と (II) においては全く無く，(III) においては標本数 n のみを無限大とする大標本漸近理論に基づくものだけしかなかった．本研究課題では，この通例が正しいかどうかを理論的に明らかにし，正しくないのであればそれを是正することで，どの情報量規準を使用すればよいか理論に基づいた判断基準を作成することにある．近年，ハードウェアの発達により，蓄積・解析できるデータの数は爆発的に増えているため，多変量モデルでの目的変数ベクトルの次元数 p が大きい高次元モデルでの解析の重要が高まっている．そのため，大標本漸近理論だけでなく，次元数 p も n と共に無限大とする高次元大標本漸近理論でも理論的な評価を行う．このように作成した判断基準は，実解析において，非常に有用なものとなると考える．

３．研究の方法

研究の方法としては，以下の通りである．

(1) Nishii (1984)では，単変量の重回帰モデルにおいて，選ばれたモデルでの当てはめ値の二乗距離に基づく予測誤差を漸近的に最小にする情報量規準は BIC のように一致性を持つ規準量であることが示されている．この結果を多変量線形回帰モデルの下で，かつモデルの良さを測る距離を KL 距離に変更しても成り立つかどうかを調べる．KL 距離に基づく予測誤差は，選ばれたモデルでの KL 距離に基づくロス関数 (KL ロス関数) の期待値として定義される．そのため，KL ロス関数が収束する先を大標本漸近理論と高次元大標本漸近理論により評価する．大標本漸近理論の下では，モデルに正規性を仮定するが真のモデルの分布はどのような分布に従っているかわからないという仮定の下で導出する．高次元大標本漸近理論の下では，真のモデルの分布も正規分布であるという仮定の下で導出する．また，小・中標本ではどの情報量規準が予測誤差を小さくするかを数値実験により確かめる．

(2) (1)と同じ正規性を仮定した多変量線形回帰モデルにおいて，真のモデルを最適なモデルとして選択する確率が標本数 n の増加と共に 1 に収束する性質である，一致性を持つための情報量規準の条件を導出する．使用する漸近理論としては，(1)と同様に，大標本漸近理論と高次元大標本漸近理論の 2 つである．得られた条件から，どのような真のモデルであっても，またどちらの漸近理論を用いたとしても，一致性を持つような情報量規準を提案する．そのような情報量規準は，標本数がある程度大き

ければ，次数 p の大小に係らず真のモデルを高い確率で選択できることが期待できる．また，小・中標本ではどの情報量規準が選択確率を高くするかを数値実験により確かめる．

４．研究成果

研究成果としては以下があげられる．

[1] 雑誌論文 1 と学会発表 2 において， KL ロス関数を最小にするモデルは真のモデルかまたは真のモデルを含んでいない過少評価モデルであることを示し，大標本漸近理論により評価すれば, 漸近的に KL ロス関数を最小にするモデルは真のモデルであることを示した．この結果から，ある程度標本数があれば，AIC のような予測 KL 距離の漸近不偏推定量ではなく，BIC などの一致性を持つ情報量規準の方が予測誤差を小さくすることがわかった．また，小・中標本では，真のモデルを含む過大評価モデルは KL ロス関数を最小にすることがないので，過少評価モデルの下でも予測 KL 距離に対するバイアスを補正した，Fujikoshi & Satoh により提案された修正 AIC (Modified AIC: MAIC) や Fujikoshi *et al.* (2005) により提案された修正拡張情報量規準 (Adjusted Extended Information Criterion: EIC_A) が予測誤差を小さくすることがわかった．

[2] 雑誌論文 3 と学会発表 4, 5 において真のモデルの分布も正規分布であるという仮定の下，一致性をもつための情報量規準の条件を導出した．大標本漸近理論により一致性を評価した場合，AIC などの予測 KL 距離の漸近不偏推定量となる情報量規準は一致性を持たず，BIC のようにモデルの複雑さに対する罰則項が標本数 n の増加に伴い無限大となるが，罰則項を n で割ったものが 0 に収束するような情報量規準が一致性を持つことがわかった．これは従来から良く知られた結果である．また，高次元大標本漸近理論により一致性を評価した場合，AIC は一致性を持ち，BIC は一致性を持たないことがあることがわかった．一致性を持つかどうかは非心パラメータ行列を n で割った行列の最大固有値の発散速度に依存しており，既存の情報量規準では，どのような真のモデルでも一致性を持つ情報量規準はないことも分かった．

[3] 雑誌論文 2 と学会発表 6, 7 において，真のモデルの分布が必ずしも正規分布に従っているかどうかわからないという仮定の下，高次元大標本漸近理論を用いて一致性を満たすための情報量規準の条件を導出した．導出のために，非心パラメータ行列を標本数 n で割った行列の最大固有値の発散速度が $O(p)$ であるという仮定を付加した．[2]で得られた条件より も若干強いものにはなったが，ほぼ同じような条件を得ることができた．

[4] 学会発表 3 において，真のモデルの分布が正規分布であるという仮定と[3]と同じ非心パラメータ行列の条件の下, KL ロス関数の収束先を高次元大標本漸近理論により評価し, KL ロス関数を漸近的に最小にするモデルが真のモデルであることを示した．この結果から, KL ロス関数を漸近的に最小にするためには少なくとも一致性を持つ情報量規準を用いる必要があることがわかった．実際には, KL ロス関数を最小にするための条件は，一致性のための条件よりも若干強いものであることもわかった．

[5] [2]で得られた結果は，非心パラメータ行列を標本数 n で割った行列の最大固有値が発散するという条件のもとでの一致性を満たすための条件であるが，その最大固有値が発散しない場合もある．学会発表 1 では，その非心パラメータ行列の条件が満たされないときでも一致性を満たす情報量規準，高次元性調整した一致性をもつ一般化情報量規準 (High-dimensionality-adjusted consistent information criterion: HCGIC) を提案した．HCGIC は大標本漸近理論の下でも，また，高次元大標本漸近理論のもとでも 一致性を持つことがわかっている．実際の HCGIC の罰則項は，(モデルのパラメータ数)×αであり，$\sqrt{p}\{\alpha+(n/p)\log(1-p/n)\}$ が無限大となるが，$(p/n)\{\alpha+(n/p)\log(1-p/n)\}$は 0 に収束するような$\alpha$である．また，数値実験により，たとえ標本数がさほど大きくなくても，HCGIC は高い確率で真のモデルを選ぶことも確かめられた．

以上の結果から，どの情報量規準を使うかどうかの判定基準として，以下が提案できる．

n	p	一致性	予測誤差
小・中	小	HCGIC	MAIC, EIC_A
大			HCGIC
小・中	大	HCGIC	MAIC, EIC_A
大			HCGIC

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 10 件）

1. Yanagihara, H., Kamo, K., Imori, S. & Yamamura, M. A study on the bias-correction effect of the AIC for selecting variables in normal multivariate linear regression models under model misspecification, REVSTAT-Statistical Journal, 査読有, 印刷中.
2. Yanagihara, H., Conditions for consistency of a log-likelihood-based information criterion

in normal multivariate linear regression models under the violation of normality assumption, Journal of the Japan Statistical Society, 45, 査読有, 2015, 21-56.
3. Yanagihara, H., Wakaki, H. & Fujikoshi, Y., A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large, The Electronic Journal of Statistics, 9, 査読有, 2015, 869-897.
4. Fukui, K., Yamamura, M. & Yanagihara, H., Comparison with RSS-based model selection criteria for selecting growth functions, FORMATH, 14, 査読有, 2015, 27-39.
5. Kamada, A., Yanagihara, H., Wakaki, H. & Fukui, K., Selecting a shrinkage parameter in structural equation modeling with a near singular covariance matrix by the GIC minimization method, Hiroshima Mathematical Journal, 44, 査読有, 2014, 315-326.
6. Hashiyama, Y., Yanagihara, H. & Fujikoshi, Y., Jackknife bias correction of the AIC for selecting variables in canonical correlation analysis under model misspecification, Linear Algebra and its Applications, 455, 査読有, 2014, 82-106.
7. Imori, S., Yanagihara, H. & Wakaki, H., Simple formula for calculating bias-corrected AIC in generalized linear models, Scandinavian Journal of Statistics, 41, 査読有, 2014, 535-555.
8. Fujikoshi, Y., Sakurai, T. & Yanagihara, H., Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression, Journal of Multivariate Analysis, 123, 査読有, 2014, 184-200.
9. Nagai, I., Fukui, K. & Yanagihara, H., Choosing the number of repetitions in the multiple plug-in optimization method for the ridge parameters in multivariate generalized ridge regression, Bulletin of Informatics and Cybernetics, 45, 査読有, 2013, 25-35.
10. Yanagihara, H., Yuan, K.-H., Fujisawa, H. & Hayashi, K., A class of cross-validatory model selection criteria, Hiroshima Mathematical Journal, 43, 査読有, 2013, 149-177.

〔学会発表〕（計 7 件）

1. Yanagihara, H. & Shimodaira, H., Consistent information criterion in normal multivariate linear regression models even under high-dimensionality, The 9th Conference of the Asian Regional Section of the International Association for Statistical Computing, December 17, 2015, Singapore (Singapore).
2. 柳原宏和, 情報量規準とバイアス補正, 日本行動計量学会岡山地域部会第 55 回研究会, 2015 年 3 月 14 日, 岡山理科大 (岡山県・岡山市).
3. Yanagihara, H., On asymptotically KL loss efficiency of a log-likelihood-based information criterion in high-dimensional normal multivariate linear regression models, The 3rd. Institute of Mathematical Statistics Asia Pacific Rim Meeting, July 1, 2014, Taipei (Taiwan).
4. Yanagihara, H., Wakaki, H. & Fujikoshi, Y., A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large, The 3rd. Institute of Mathematical Statistics Asia Pacific Rim Meeting, June 30, 2014, Taipei (Taiwan).
5. 若木宏文・柳原宏和・藤越康祝, A consistency property of the AIC for multivariate linear models when the dimension, the sample size and the number of explanation variables are large, 2013 年度統計関連学会連合大会, 2013 年 9 月 10 日, 大阪大学 (大阪府・吹田市).
6. 柳原宏和, Effects of nonnormality on a consistency property of several information criteria for multivariate linear regression models when the dimension and the sample size are large, 2013 年度統計関連学会連合大会, 2013 年 9 月 10 日, 大阪大学 (大阪府・吹田市).
7. Yanagihara, H., Conditions for consistency of a log-likelihood-based information criterion in high-dimensional multivariate linear regression models under the violation of normality assumption, 22nd. International Workshop on Matrices and Statistics, August 15, 2013, Toronto (Canada).

６．研究組織

(1)研究代表者

柳原　宏和（YANAGIHARA HIROKAZU）
広島大学・大学院理学研究科・准教授
研究者番号：70342615

(3)連携研究者

藤澤　洋徳（FUJISAWA HIRONORI）
統計数理研究所・数理・推論研究系・教授
研究者番号：00301177

二宮　嘉行（NINOMIYA YOSHIYUKI）
九州大学・マス フォア インダストリ研究所・准教授
研究者番号：50343330